International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Evaluation of Biomarkers Dataset for Patients Consumed With and Without Antioxidant: An Approach by Machine Learning Algorithms

Sk Ziaur Rahaman

PhD, Rahaman Clinic, Y-211/B Panchpara Road, Garden Reach, Kolkata – 700018, India
Corresponding Author Email: *dr_ziaur[at]rediffmail.com*

Abstract: ***Considerable attention has been paid to assessing the impact of antioxidants on different biological markers due to their increasing application in the prevention and treatment of illnesses brought by oxidative stress. Complex, nonlinear correlations between several biomarkers cannot be identified using conventional statistical techniques. Machine learning (ML) algorithms offer advanced analytical techniques for recognizing latent patterns, categorizing patient groups, and forecasting treatment outcomes. This research seeks to use ML techniques to compare biomarker data from individuals who took antioxidants with those who did not. In the present study, biomarker dataset mining through ML algorithms viz. Bayes Network (BN), NaiveBayes (NB), Logistic Regression (LR), K-nearest neighbour (IBK), Instance-based classifier (K*) and LogitBoost (LB) were studied by using WEKA (Waikato Environment for Knowledge Analysis) tool (version, 3.8.5). The prediction of accuracy as per effect class (normal and abnormal) values related to statistical interpretations on MCC, ROC and PRC in which the highest values were predicted in K* and LB followed by BN, NB, IBK and LR. By improving knowledge of antioxidant effectiveness and promoting clinical decision-making, the findings have the potential to further personalized medicine.***

Keywords: Algorithms, Biomarkers, Dataset, Machine learning, Prediction accuracy

1. Introduction

An imbalance between the creation of reactive oxygen species (ROS) and the body's antioxidant defense systems causes oxidative stress. [1] It is essential to the development of aging, neurodegenerative disorders, cancer, diabetes, and cardiovascular illnesses. Free radicals are neutralized and oxidative damage is decreased via antioxidants, which are either acquired through food or supplements.[2] Many studies evaluated the common biomarkers for determining oxidative stress and antioxidant properties prevented the disease after inhibiting these markers. [3-5]

Big datasets on biomarker are now accessible due to advancements in biomedical data gathering. These variables' complicated interactions may not be accurately predicted by traditional statistical analyses. Multidimensional data may be processed, key features identified, and extremely accurate predictive models produced by ML algorithms.[5] Some studies have been conducted for big data mining on biomedical sciences.[6-7]

In this regard, this study proposes a ML-based framework to evaluate biomarker datasets from patients who consumed antioxidants and those who did not, aiming to identify potential biomarkers and to compare the predictive performance of different ML algorithms.

2. Materials and Methods

In the present study, biomarker dataset mining through ML algorithms viz. Bayes Network (BN), NaiveBayes (NB), Logistic Regression (LR), K-nearest neighbour (IBK), Instance-based classifier (K*) and LogitBoost (LB) were studied by using WEKA (Waikato Environment for Knowledge Analysis) tool (version, 3.8.5) developed by Frank et al. [8] The WEKA explorer was developed with data pre-processing, classification, regression, and association rules.[9] After preprocessing of dataset classification was performed as 10-fold cross validation (CV) and the predictive accuracy was determined for each algorithm. Different attributes such as FBS, PPBS, HbA1c, MDA, IL-6 and CRP parameters and the effect (normal and abnormal values) were used.

According to correctly and incorrectly classified instances, Kappa (K) statistics, mean absolute error (MAE), and root mean squared error (RMSE), the model accuracy of the aforementioned ML algorithm classifications was evaluated. According to Bouckaert et al.,[9] the WEKA tool was utilized to retrieve and consider the modelling result summary. The statistical parameters were "Matthew's correlation coefficient (MCC), receiver operating characteristic (ROC), and precision-recall curve (PRC)", respectively.

3. Results

Table 1 describes the summary results of correctly and incorrectly classified instances of studied algorithms. In the case of algorithm classification, the highest values were obtained in K* and LB followed by BN, NB, IBK and LR as per 10-fold CV.

Table 1: Summary results of different models (correctly and incorrectly classified instances)

Classifier models	CCI (%)	ICI (%)
BN	98.33	1.67
NB	98.33	1.67
LR	95.00	5.00
IBK	98.33	1.67
K*	100.00	0.00
LB	100.00	0.00

CCI = Correctly classified instances, ICI = Incorrectly classified instances, BN = Bayes Network, NB = NaiveBayes, LR = Logistic Regression, IBK = K-nearest neighbour, K* = Instance-based classifier, LB = LogitBoost

Table 2 describes the model summary results of Kappa statistic (KS), mean absolute error (MAE) and root mean squared error (RMSE) of studied models related to 10-fold CV. In case of prediction accuracy of the class of KS values, the highest values were observed in K* and LB followed by BN, NB, IBK and LR while lower values of MAE and RMSE in the same way.

Table 2: Model summary (Kappa statistic, mean absolute error and root mean squared error) results

Classifier models	KS	MAE	RMSE
BN	0.97	0.02	0.13
NB	0.97	0.02	0.14
LR	0.90	0.05	0.22
IBK	0.97	0.03	0.13
K*	1.00	0.009	0.05
LB	1.00	0.00	0.00

KS = Kappa statistic, MAE = Mean absolute error; RMSE = Root mean squared error, BN = Bayes Network, NB = NaiveBayes, LR = Logistic Regression, IBK = K-nearest neighbour, K* = Instance-based classifier, LB = LogitBoost

Table 3 represents the prediction of accuracy as per effect class (normal and abnormal) values related to statistical interpretations on MCC, ROC and PRC in which the highest values were predicted in K* and LB followed by BN, NB, IBK and LR.

Table 3: Summarized statistical results for studied algorithms

Classifier models	MCC	ROC	PRC	Class
BN	0.97	0.99	0.99	Normal
	0.97	0.99	0.99	Abnormal
NB	0.96	0.99	0.99	Normal
	0.96	0.99	0.99	Abnormal
LR	0.90	0.99	0.99	Normal
	0.90	0.99	0.99	Abnormal
IBK	0.97	0.98	0.97	Normal
	0.97	0.98	0.98	Abnormal
K*	1.00	1.00	1.00	Normal
	1.00	1.00	1.00	Abnormal
LB	1.00	1.00	1.00	Normal
	1.00	1.00	1.00	Abnormal

BN = Bayes Network, NB = NaiveBayes, LR = Logistic Regression, IBK = K-nearest neighbour, K* = Instance-based classifier, LB = LogitBoost

4. Discussion

Interestingly, the WEKA tool evaluates easily ML classification algorithms and the prediction accuracies of 10-fold CV for the current dataset related to classified instances, error rate, and kappa statistics and other statistical analysis data can be provided easily within this tool.

In the present study, MCC, ROC and PRC data for normal and abnormal effect class in which the K* and LB followed by BN, NB, IBK and LR, which are supported by different past studies in biomedical sciences. [6-7]

In earlier study, these datasets were used with ML Tree algorithms, but the predictive accuracy was similar for Fast decision tree learner (REPTree) and SimpleCart (SC) algorithms.[10]

5. Conclusions

For evaluating biomarker data from patients with and without antioxidant consumption, machine learning algorithms offer a strong framework. Precision medicine methods can be supported, patient groups can be categorized precisely, and important biomarkers can be identified with the help of ML by combining cutting-edge feature selection and predictive modelling techniques. In this study, it was predicted that antioxidant consumption patients were found a declining trend for the risk factors related to biomarkers due to antioxidants consumptions when compared to without consumptions group.

Conflict of interest
Authors declare no conflict of interest.

Funding
This is non-funded project.

References

[1] Birben E, Sahiner UM, Sackesen C, Erzurum S, Kalayci O. Oxidative stress and antioxidant defense. World Allergy Organ J. 2012;5(1):9-19.
[2] Chandimali N, Bak SG, Park EH, Lim HJ, Won YS, Kim EK, et al. Free radicals and their impact on health and antioxidant defenses: a review. Cell Death Discov. 2025;11(1):19.
[3] Valko M, Leibfritz D, Moncol J, Cronin MT, Mazur M, Telser J. Free radicals and antioxidants in normal physiological functions and human disease. Int J Biochem Cell Biol. 2007;39(1):44-84.
[4] Forman HJ, Zhang H. Targeting oxidative stress in disease: promise and limitations of antioxidant therapy. Nat Rev Drug Discov. 2021;20(9):689-709.
[5] Cordiano R, Di Gioacchino M, Mangifesta R, Panzera C, Gangemi S, Minciullo PL. Malondialdehyde as a potential oxidative stress marker for allergy-oriented diseases: an update. Molecules. 2023;28(16):5979.
[6] Ng S, Masarone S, Watson D, Barnes MR. The benefits and pitfalls of machine learning for biomarker discovery. Cell Tissue Res. 2023; 394:17-31.
[7] Asaoka R, Hirasawa K, Iwase A, Fujino Y, Murata H, Shoji N, et al. Validating the usefulness of the “random forests” classifier to diagnose early glaucoma with optical coherence tomography. American Journal of Ophthalmology. 2017; 174: 95-103.
[8] Bavikadi D, Agarwal A, Ganta S, Chung Y, Song L, Qiu J, Shakarian P. Machine learning driven biomarker selection for medical diagnosis. PLoS One. 2025;20(6):e0322620.
[9] Frank E, Hall MA, Witten IH, The WEKA workbench, Online appendix for data mining: Practical machine learning tools and techniques. Morgan Kaufmann, Fourth Edition, 2016.
[10] Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, et al. WEKA manual for version 3-8-5.

University of Waikato, Hamilton, New Zealand, December 21, 2020.

[11] Rahaman SZ, Banerjee SK. The risk factors in type-2 diabetes mellitus with and without antioxidant consumed patients: prediction accuracy through machine learning algorithms. Journal Of Electronics Information Technology Science and Management. 2022;12(10):141-151.